Técnicas Econométricas Utilizadas

1. **Equação Minceriana de Salário (Renda do Trabalho)**

A equação minceriana de salários serve de base a uma vasta literatura empírica de economia do trabalho. O modelo salarial de Jacob Mincer é o arcabouço utilizado para estimar retornos da educação, entre outras variáveis determinantes da renda do trabalho. Mincer concebeu uma equação para rendimentos que seria dependente de fatores explicativos associados à escolaridade e à experiência, além de possivelmente outros atributos, como sexo, por exemplo.

Essa equação é a base da economia do trabalho em particular no que tange aos efeitos da educação. Sua estimação já motivou centenas de estudos, que tentam incorporar diferentes custos educacionais, como impostos, mensalidades, custos de oportunidades, material didático, assim como a incerteza e a expectativa dos agentes presentes nas decisões, o progresso tecnológico, não linearidades na escolaridade etc. Identificando os custos da educação e os rendimentos do trabalho, viabilizou o cálculo da taxa interna de retorno da educação, que é a taxa de desconto que equaliza o custo e o ganho esperado de se investir em educação –- a taxa de retorno da educação, que deve ser comparada com a taxa de juros de mercado para determinar a quantidade ótima de investimento em capital humano. A equação de Mincer também é usada para analisar a relação entre crescimento e nível de escolaridade de uma sociedade, além dos determinantes da desigualdade. Uma de suas grandes virtudes é incorporar em uma só equação dois conceitos econômicos distintos:

a) uma equação de preço revelando quanto o mercado de trabalho está disposto a pagar por atributos produtivos como educação e experiência e

(b) a taxa de retorno da educação, que deve ser comparada com a taxa de juros de mercado para determinar a quantidade ótima de investimento em capital humano.

Modelo da Regressão

O modelo econométrico de regressão típico decorrente da equação minceriana é:

$$\ln w = \beta 0 + \beta 1\ educ + \beta 2\ exp + \beta 3\ exp^2 + \gamma' x + \epsilon$$

onde
w é o rendimento do trabalho recebido pelo indivíduo;
educ é a sua escolaridade, geralmente medida por anos de estudo;
exp é sua experiência, geralmente aproximada pelo idade do indivíduo;
x é um vetor de características observáveis do indivíduo, como raça, gênero, região; e
ϵ é um erro estocástico.

Este é um modelo de regressão no formato log-nível, isto é, a variável dependente – o salário – está em formato logaritmo e a variável independente mais relevante - a escolaridade – está em nível. Portanto, o coeficiente $\beta 1$ mede quanto um ano a mais de escolaridade causa de variação proporcional no salário do indivíduo. Por exemplo, se $\beta 1$ é estimado em 0,18, isso quer dizer que cada ano a mais de estudo está relacionado, em média, com um aumento de rendimento de 18%.

Derivando, encontramos que $(\partial \ln w / \partial\ educ) = \beta 1$

Por outro lado, pela regra da cadeia, tem-se que:

$$(\partial \ln w / \partial\ educ) = (\partial w / \partial\ educ)\ (1 / w) = (\partial w / \partial\ educ) / w)$$

Logo, $\beta 1 = (\partial w / \partial\ educ) / w$, correspondendo avariação percentual do salário decorrente de cada acréscimo unitário de ano de estudo.

Principais dificuldades

Entre os principais problemas das equações mincerianas de determinação de salário estão:

1) Viés de não habilidade não-observável: a habilidade afeta positivamente tanto a escolaridade quanto o salário. Portanto, na verdade parte do retorno à educação verificada

se deve na verdade a uma maior habilidade do indivíduo, que por si só gera aumento de salário, e não a educação propriamente dita. Esse viés vai na direção de uma superestimação dos retornos à educação.

2) Erro de medida: pessoas descrevem sem exatidão sua escolaridade. Como geralmente elas reportam o nível de escolaridade correto ou acima do correto, arredondando para cima um ano ou um ciclo inteiro, o retorno encontrado vai se encontrar abaixo do correto. Logo, esse erro vai na direção de uma subestimação dos retornos à educação.

Uma vantagem é que esses dois principais problemas vão cada um em uma direção, o que faz com que se compensem em alguma medida. Outros pontos sensíveis a serem destacados são:

3) Em lugares nos quais indivíduos mais educados trabalham mais do que indivíduos menos educados, parte dos diferenciais de salário podem estar refletindo mais horas trabalhadas, e vice-versa.
4) Vários benefícios da escolaridade não são considerados no cálculo desses retornos, como seu retorno nas dimensões políticas, psicológicas, filosóficas e inúmeras outras dimensões não monetárias.

Como o mercado determina o retorno à educação

Na prática, o retorno a educação pode ser entendido como o preço que o mercado de trabalho, regido pelas leis de oferta e demanda, determina para o atributo educação.

Observamos o equivalente a uma corrida entre a oferta de qualificação da mão-de-obra, proporcionada por uma expansão da educação e entre a demanda por mão de obra qualificada, advinda do progresso tecnológico. É justamente a tensão essa tensão entre demanda e oferta do

atributo educação que define seu preço, na forma do retorno à educação. Langoni encontra, por exemplo, para o caso brasileiro na década de 70, que a educação deveria se expandir a uma taxa de 1,23% ao ano para ganhar a corrida contra o progresso tecnológico, impedindo que os retornos se elevassem ainda mais, o que aumentaria mais a desigualdade.

2. Regressão logística

Definimos variáveis categóricas como aquelas variáveis que podem ser mensurados usando apenas um número limitado de valores ou categorias. Esta definição distingue variáveis categóricas de variáveis contínuas, as quais, em princípio, podem assumir um número infinito de valores. Muitas variáveis de interesse para cientistas sociais são claramente categóricas, entre as quais podemos destacar raça, gênero, estado civil, emprego, nascimento e morte.

O tipo de regressão utilizado nos simuladores, assim como para determinar as diferenças-em-diferenças, é o da regressão logística, método empregado para estudar variáveis dummy -- aquelas compostas apenas por duas opções de eventos, como “sim” ou “não”. Por exemplo:

Seja Y uma variável aleatória *dummy* definida como:

$$Y = \begin{cases} 1 \text{ se a pessoa estava ocupada} \\ 0 \text{ se a pessoa não estava ocupada} \end{cases}$$

Cada Y_i tem distribuição de Bernoulli, cuja função de distribuição de probabilidade é dada por:

$$P(y \mid p) = p^{y}(1 - p)^{1-y}$$

Onde: y identifica o evento ocorrido e p é a probabilidade de sucesso de ocorrência do evento.

Como se trata de uma seqüência de eventos com distribuição de Bernoulli, a soma do número de sucessos ou fracassos neste experimento tem distribuição binomial de parâmetros n (número de

observações) e p (probabilidade de sucesso). A função de distribuição de probabilidade da binomial é dada por:

$$P(y \mid n, p) = \binom{n}{y} p^{y} (1 - p)^{1-y}$$

A transformação logística pode ser interpretada como o logaritmo da razão de probabilidades sucesso versus fracasso, no qual a regressão logística nos dá uma idéia do retorno de uma pessoa obter ocupação, dado o efeito de algumas variáveis explicativas que serão introduzidas mais à frente, em particular a educação profissional.

A função de ligação deste modelo linear generalizado é dada pela seguinte equação:

$$\eta_i = \log\left(\frac{p_i}{1 - p_i}\right) = \sum_{k=0}^{K} \beta_k x_{ik}$$

onde a probabilidade pi é dada por:

$$p_i = \frac{\exp\left(\sum_{k=0}^{K} \beta_k x_{ik}\right)}{1 + \exp\left(\sum_{k=0}^{K} \beta_k x_{ik}\right)}$$

Os modelos utilizados aqui têm como objetivo identificar as variáveis relacionadas com as características de interesse (variável resposta). Ao realizar o ajuste do modelo, deseja-se encontrar, e identificar, quais são os fatores importantes que melhor descrevem o comportamento/variação das características de interesse.

O modelo linear generalizado aqui utilizado é definido por uma distribuição de probabilidade para a variável resposta, um conjunto de variáveis independentes (fatores explicativos) que compõem

o previsor linear do modelo, e uma função de ligação entre a média da variável resposta e o referido previsor linear.

3) Modelo Logit Multinomial[1]

Muitos estudos de relevância social são mensurados através de variáveis qualitativas não ordenadas. Por exemplo, sociólogos e economistas estão interessados na composição da força de trabalho (empregados, desempregados); cientistas políticos em afiliações partidárias (direita, esquerda); geógrafos e demógrafos nas regiões de residência (Nordeste, Norte, Sul, etc.).

É um dos muitos métodos utilizados para analisar variáveis de resposta categórica não ordenada (nominal) nas pesquisas de ciências sociais. Podemos citar algumas razões para esta popularidade: tal modelo é uma generalização do modelo logit binomial; é equivalente para o modelo log-linear com dados agrupados e; estão disponíveis no mercado de vários softwares estatísticos para o ajuste destes modelos.

Quando dizemos que uma variável é não ordenada, dizemos que cada categoria é única em comparação com outras categorias.

Para o resultado da variável (y) com J categorias (j=1, ..., J), vejamos a diferença da j-ésima (j>1) categoria com a primeira ou a categoria base, derivando a base logit para a j-ésima categoria.

$$B_j = \log\left[\frac{P(y=j)}{P(y=1)}\right] = \log\left(\frac{p_j}{p_1}\right), j = 2,...,J \longrightarrow (1)$$

[1] Esta seção baseia-se no livro Statistical Methods for Categorical Data Analysis – Daniel A Powers, Yu Xie – capítulo 7.

Onde pj e p1 denotam as probabilidades da j-ésima e primeira categoria. A escolha do uso da primeira categoria como base foi arbitrária.

Alguma outra categoria poderia ser usada como base. Na transformação da estrutura, nós podemos retornar a base do logit especificado na Eq. (1) como função linear de x. Contanto, é necessário especificar a categoria de contraste (isto é j) como também a categoria base (1 neste caso) quando modelamos resultados qualitativos não ordenados. Existe J-1 bases não redundantes para resultados de variáveis com J categorias.

Agora consideramos o caso de termos apenas uma variável independente x com um número limitado de categorias (x=1,...I). Este caso é equivalente a tabela de contigência, cada valor de x (x=i), a base é:

$$\log\left[\frac{P(y=j/x=i)}{P(y=1/x=i)}\right]=\log\left[\frac{p_{ij}}{p_{i1}}\right]=B_{ij}\longrightarrow(2)$$

Considerando neste contexto temos especificado um modelo saturado, a estimação da Eq (2) pode ser obtida como:

$$\log\left[\frac{F_{ij}}{Fi1}\right]=\log\left[\frac{f_{ij}}{f_{i1}}\right],\longrightarrow(3)$$

onde fij e Fij, são as frequências observada e esperada na i-ésima linha e j-ésima coluna para a classificação da tabela X x Y. Nós podemos facilmente rescrever o resultado na forma de Modelo Linear Generalizado:

$$B_{ij}=\sum_{i=1}^{I}\log\left(\frac{F_{ij}}{F_{i1}}\right).I(x=i)\longrightarrow(4)$$

onde I(.) é a função indicadora, I=1 se verdadeira, 0, caso contrário. Com variável dummy codificando e a primeira categoria como referência, Eq. (4) é usualmente escrita como:

$$B_{ij} = \alpha_j \sum_{i=1}^{I} \beta_{ij} . I(x = i), x > 1, \longrightarrow (5)$$

onde αj é a base para x=1, e βij é a diferença na base entre x=i e x=1, Nesse caso simples, αj e βij podem ser estimados separadamente para todo i e j. Estimações simultâneas resultarão num modelo equivalente neste caso. Para outros modelos do que o modelo saturado, separar e estimar simultaneamente em geral gera resultados diferentes.

Modelo Logit Multinomial padrão

Vejamos agora a uma situação mais geral com dados individuais e mudanças na notação dado que i agora represente o i-ésimo indivíduo. Seja yi uma variável com resultados politômicos com categorias codificadas por 1, ..., J. Associando com cada categoria é uma probabilidade de resposta, (Pi1, Pi2,...PiJ) representam a chance do i-ésimo respondente numa categoria particular.

Como no caso de resultados binários, assumimos a presença de um vetor que mede características dos respondentes, xi (incluindo 1 como o primeiro elemento), como preditores das probabilidades respondente.

Utilizando a notação da função índice, a resposta para a probabilidade depende de transformações não lineares da função linear Xiβij = ∑k=0 βjk xik , onde k é o número de preditores (na notação, o primeiro parâmetro B0 é o termo de intercepto, o mesmo alfa da eq. 8). É importante notar que, os casos para modelo binomial logit, os parâmetros no modelo multinomial logit apresentam vários resultados categóricos.

O modelo multinomial logit pode ser visto como uma extensão do modelo binário logit, expresso pela eq. 2 e 3, situações em que o resultado das variáveis tem múltiplas categorias não ordenadas. Por exemplo, no caso de três categorias (J=3), nós podemos escrever as probabilidades:

$$\Pr(y_i = 1/x_i) = P_{i1} = \frac{1}{1+\exp(x_i'\beta_{12})+\exp(x_i'\beta_{13})},$$

$$\Pr(y_i = 2/x_i) = P_{i1} = \frac{\exp(x_i'\beta_2)}{1+\exp(x_i'\beta_{22})+\exp(x_i'\beta_{23})}$$

$$\Pr(y_i = 3/x_i) = P_{i1} = \frac{\exp(x_i'\beta_3)}{1+\exp(x_i'\beta_{32})+\exp(x_i'\beta_{33})},$$

Onde β2 e β3 denotam os efeitos das covariáveis especificadas para a segunda e terceira categorias de resposta com a primeira categoria usada como referência. Note que a equação Pi1 é derivada do contraste entre a soma das três probabilidades que é 1. Isto é, Pi1=1-(Pi2+Pi3), onde

$$P_{i1} = \frac{\eta_{i1}}{\eta_{i1}+\eta_{i2}+\eta_{i3}},$$

$$P_{i2} = \frac{\eta_{i2}}{\eta_{i1}+\eta_{i2}+\eta_{i3}}, \longrightarrow (10)$$

$$P_{i3} = \frac{\eta_{i3}}{\eta_{i1}+\eta_{i2}+\eta_{i3}},$$

yi = 1 define a base.

As probabilidades da equação acima podem ser expressas em termos da função exponencial dos termos lineares ηij=exp(x'iβj):

Estimação

A estimação é obtida iterativamente usando máxima verossimilhança. É conveniente definir um conjunto de J variáveis dummy: dij=1 se yi=j e 0 caso contrário. Este resultado em um e apenas um dij=1 para cada observação. O log da verossimilhança é:

$$\log L = \sum_{i=1}^{n} \sum_{j=1}^{J} d_{ij} \log Pij \longrightarrow (13)$$

Interpretando os resultados de um Modelo Logit Multinomial - Vantagem e Razão de vantagem

Uma importante parte do modelo multinomial somente como elas são em respostas binárias e modelos loglineares. Na estrutura modelo multinomial logit, a vantagem entre categorias j e 1 é

$$\frac{P_{ij}}{P_{i1}} = \frac{\eta_{ij}}{\eta_{i1}} = \exp(x_i^{'} \beta_j) \longrightarrow j = 2,...J \longrightarrow (14)$$

dada por i simplesmente:

O log da vantagem, ou logit, está na função linear de xi:

Razão de vantagens

Às vezes temos interesse em conhecer a vantagem do sucesso de um grupo, mais especificamente se tem conta. Um exemplo para esse caso seria a seguinte questão: será que a vantagem de uma pessoa com alta escolaridade ter acesso a conta é e o quanto é maior que a de uma de baixa escolaridade? A razão de vantagens seria uma boa forma de medir isso.

A razão de vantagens é dada pela seguinte relação:

$$\theta = \left(\frac{p_1}{1 - p_1} \right) \Big/ \left(\frac{p_2}{1 - p_2} \right)$$

onde p_1 e p_2 são as probabilidades de sucesso dos grupos 1 e 2, respectivamente.

Assim, percebe-se que a razão de vantagens, ou razão condicional, difere da probabilidade. Exemplificando-se novamente: se um cavalo tem 50% de probabilidade de vencer uma corrida, sua razão condicional é de 1 em relação aos outros cavalos, isto é, sua chance de vencer é de um para um. O conceito de razão condicional é de extrema importância para a compreensão deste trabalho, pois nos indicará se a variável gerada por diferenças-em-diferenças aumentou ou diminuiu a chance de sucesso em relação à variável estudada.

Seleção de Variáveis

Para selecionar o modelo utilizou-se a PROC GENMOD do SAS (maiores detalhes em www.sas.com). Os modelos finais foram selecionados passo a passo, após agrupamento de níveis dos fatores com base na estatística de Wald, incluindo-se em cada passo as interações que produziam maior decréscimo da Deviance, considerando o teste da razão.

Os modelos finais foram selecionados passo a passo, após agrupamento de níveis dos fatores com base na estatística de Wald, incluindo-se em cada passo as interações que produziam maior decréscimo da Deviance, considerando o teste da razão.

Estimador de diferença em diferença

Em Ciências Sociais, muitas pesquisas são feitas analisando os chamados experimentos. Para analisar um experimento natural sempre é preciso ter um grupo de controle, isto é, um grupo que não foi afetado pela mudança, e um grupo de tratamento, que foi afetado pelo evento, ambos com características semelhantes. Para estudar as diferenças entre os dois grupos são necessários dados

de antes e de depois do evento para os dois grupos. Assim, a amostra está dividida em quatro grupos: o grupo de controle de antes da mudança, o grupo de controle de depois da mudança, o grupo de tratamento de antes da mudança e o grupo de tratamento de depois da mudança.

A diferença entre a diferença verificada entre os dois períodos, entre cada um dos grupos é a diferença em diferença, representada com a seguinte equação:

$g3 = (y2,b – y2,a) – (y1,b – y1,a)$

Onde cada y representa a média da variável estudada para cada ano e grupo, com o número subscrito representando o período da amostra (1 para antes da mudança e 2 para depois da mudança) e a letra representando o grupo ao qual o dado pertence (a para o grupo de controle e b para o grupo de tratamento). E g3 é a estimativa a partir da diferença em diferença. Uma vez obtido o g3, determina-se o impacto do experimento natural sobre a variável que se quer explicar.